Data Spectroscopy: Learning Mixture Models using Eigenspaces of Convolution Operators

Tao Shi TAOSHI@STAT.OSU.EDU
Department of Statistics, Ohio State University

Mikhail Belkin MBELKIN@CSE.OSU.EDU
Department of Computer Science and Engineering, Ohio State University

Bin Yu BINYU@STAT.BERKELEY.EDU
Department of Statistics, University of California Berkeley

Abstract

In this paper we develop a spectral framework for estimating mixture distributions, specifically Gaussian mixture models. In physics, spectroscopy is often used for the identification of substances through their spectrum. Treating a kernel function $K(x, y)$ as "light" and the sampled data as "substance", the spectrum of their interaction (eigenvalues and eigenvectors of the kernel matrix K) unveils certain aspects of the underlying parametric distribution p, such as the parameters of a Gaussian mixture. Our approach extends the intuitions and analyses underlying the existing spectral techniques, such as spectral clustering and Kernel Principal Components Analysis (KPCA).

We construct algorithms to estimate parameters of Gaussian mixture models, including the number of mixture components, their means and covariance matrices, which are important in many practical applications. We provide a theoretical framework and show encouraging experimental results.

1. Introduction

Gaussian mixture models are a powerful tool for various tasks of data analysis, modeling and exploration. The basic problem is to estimate the parameters of a Gaussian mixture distribution $p(x) = \sum_{g=1}^{G} \pi^g p^g(x)$, from sampled data $x_1, \ldots, x_n \in \mathbb{R}^d$, where the mixture component $p^g = N(\mu^g, \Sigma^g)$ has the mean μ^g and the covariance matrix Σ^g, $g = 1, \ldots, G$. Gaussian mixture models are used in a broad range of scientific and engineering applications, including computer vision, speech recognition, and many other areas.

However, effectiveness of modeling hinges on choosing the right parameters for the mixture distribution. The problem of parameter selection for mixture models has a long history, going back to the work of (Pearson, 1894, [9]), who introduced the Method of Moments and applied it to the study of a population of Naples crabs, deducing the existence of two subspecies within the population.

The most commonly used method for parameter estimation is *Maximum Likelihood Estimation* (MLE), which suggests choosing the parameters in a way that maximizes the likelihood of the observed data, given a model. In modern practice this is most commonly done through the iterative optimization technique known as Expectation Maximization (EM) algorithm ([3]), which is typically initialized using k-means clustering. Recently significant progress on understanding theoretical issues surrounding learning mixture distributions and EM has been made in theoretical computer science, e.g., [2, 4].

Another set of methods for inferring mixture distribution is based on the Bayesian inference, which is done using a prior distribution on the parameters of the model. In recent literature ([7]) the Dirichlet process mixture models were used to produce posterior distribution for parameters of a mixture model. The inference procedure involves applying Markov Chain Monte-Carlo to draw samples from the posterior distribution.

Appearing in *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008. Copyright 2008 by the author(s)/owner(s).

In this paper we propose a new method for estimating parameters of a mixture distribution, which is closely related to non-parametric spectral methods, such as spectral clustering (e.g., [8]) and Kernel Principal Components Analysis [11]. Those methods, as well as certain methods in manifold learning (e.g., [1]), construct a kernel matrix or a graph Laplacian matrix associated to a data set. The eigenvectors and eigenvalues of that matrix can then be used to study the structure of the data set. For example, in spectral clustering the presence of a small non-zero eigenvalue indicates the presence of clusters, while the corresponding eigenvector shows how the data set should be split. In particular, we note the work [12] where the authors analyze dependence of spectra on the input density distribution in the context of classification and argue that lower eigenfunctions can be truncated without sacrificing classification accuracy. We will develop the intuitions and analyses underlying these methods and take them a step further by offering a framework, which can be applied to analyzing parametric families, in particular a mixture of Gaussian distributions.

We would like to study mixture distributions by building explicit connections between their parameters and spectral properties of the corresponding kernel matrices. More specifically, we construct a family of probability-dependent operators and build estimators by matching eigenvalues and eigenfunctions of the operator associated to a probability distribution to those of the matrix associated to a data sample. Thus given a mixture distribution $p(x) = \sum_{g=1}^{G} \pi^g p^g(x)$, we use a Gaussian kernel $K(x,y) = e^{-\frac{\|x-y\|^2}{2\omega^2}}$ to construct the integral operator

$$\mathcal{G}_p^\omega f(y) = \int e^{-\frac{\|x-y\|^2}{2\omega^2}} f(x)\, p(x) dx$$

which will be the principal object of this paper. Our framework will rely on three key observations about the spectral properties of this operator and its connection to the sampled data.

Observation 1. (Single component) For the Gaussian distribution $p = N(\mu, \Sigma)$, we can analytically express eigenfunctions and eigenvalues of $\mathcal{G}_p^\omega$ in terms of the mean μ and the covariance Σ. This will allows us to reverse this dependence and explicitly express μ and Σ in terms of the spectral properties of $\mathcal{G}_p^\omega$.

Observation 2. (Mixture of components)

Let p be a mixture distribution $p(x) = \sum_{g=1}^{G} \pi^g p^g(x)$. Note that by linearity

$$\mathcal{G}_p^\omega f(y) = \sum_{g=1}^{G} \pi^g \int e^{-\frac{\|x-y\|^2}{2\omega^2}} f(x)\, p^g(x) dx$$

$$= \sum_{g=1}^{G} \pi^g \mathcal{G}_{p^g}^\omega f(y)$$

It can be seen (Theorem 1) that given enough separation between the mixture components, top eigenfunctions of the individual components $\mathcal{G}_{p^g}^\omega$ are approximated by top eigenfunctions of $\mathcal{G}_p^\omega$. That will allow us to connect eigenfunctions/eigenvalues of the mixture to eigenfunctions/eigenvalues of the individual components. A specific example of this is given in Fig. 2, which will be discussed in detail in Section 4.

Observation 3. (Estimation from data) The eigenfunctions and eigenvalues of $\mathcal{G}_p^\omega$ can be approximated given data sampled from $p(x)$ by eigenvectors and eigenvalues of empirical kernel matrices.

To highlight the effectiveness of our methodology consider the distribution in Fig. 1, where the density given by a mixture of two normal distributions $p = 0.9\, N(-3, 1^2) + 0.1\, N(0, 0.3^2)$ and a histogram obtained by sampling 1000 points are shown. From the Table 1, we see that the spectroscopic estimator has no difficulty providing reasonably accurate estimates for the mixing coefficients π^1, π^2, means μ^1, μ^2 and variances σ^1, σ^2 for each component, despite the fact that the mixture is unbalanced. We also see that these estimates can be further improved by using the spectroscopic estimate to initialize EM.

We note that, while EM is a computationally efficient and algorithmically attractive method, it is a local optimization procedure and the quality of the achieved maximum and accuracy of the resulting estimate are sensitive to initialization (see, e.g., [10]). If the initial value happens to be close to the global maximum, fast convergence can be guaranteed. However, finding such "lucky" regions of the parameter space may be nontrivial. To emphasize that point, consider the bottom two rows of Table 1, where the results of k-means clustering ($k = 2$) and EM initialized by k-means are shown. We see that k-means consistently provides a poor starting point as the energy minimizing configuration splits the large component, ignoring the small one. EM, initialized with k-means, stays at a local maximum and cannot provide an accurate estimate for the mixture. On the other hand, EM initialized with our method, converges to the correct solution.

We should note that our method requires sufficient separation between the components to provide accurate results. However there does not exist a computationally feasible method for estimating parameters of a mixture distribution in several dimensions without a separation assumption.

The rest of the paper is structured as follows: in Sec-

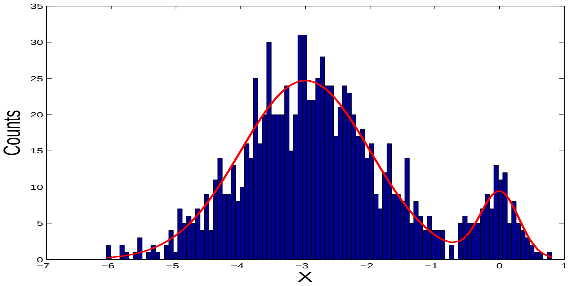

Figure 1. Histogram of 1000 data points sampled from $0.9N(-3, 1^2) + 0.1N(0, 0.3^2)$ and the distribution (red line).

True Parameters	$\pi^1 = 0.9$	$\pi^2 = 0.1$	$\mu^1 = -3$	$\mu^2 = 0$	$\sigma^1 = 1$	$\sigma^2 = 0.3$
Spectroscopic Estimator	0.86 (0.01)	0.14 (0.01)	-2.98 (0.23)	-0.02 (0.08)	1.12 (0.54)	0.34 (0.10)
EM [**SE** initialization]	0.90 (0.01)	0.10 (0.01)	-3.01 (0.04)	0.00 (0.03)	1.00 (0.03)	0.30 (0.02)
k-means [random samples]	0.68 (0.03)	0.32 (0.03)	-3.42 (0.06)	-1.17 (0.16)	0.74 (0.03)	0.90 (0.03)
EM [k-means initialization]	0.78 (0.07)	0.22 (0.07)	-3.17 (0.09)	-0.93 (0.56)	0.92 (0.05)	0.95 (0.39)

Table 1. Mixture Gaussian parameters and corresponding estimators from Spectroscopic Estimation, and EM (initialized by SE), k-means (random initialization) and EM (initialized by k-means). The mean and the (standard deviation) of each estimator over 50 runs are shown.

tion 2, we describe our approach in the simplest setting of a one-dimensional component in $\mathbb{R}$. In Section 3, we analyze a single component in $\mathbb{R}^d$, in Section 4, we deal with a general case of a mixture distribution and state a basic theoretical result for the mixture. In section 5, we show some experimental results on a simulated mixture distribution with three components in $\mathbb{R}^5$ and show some experimental results on the USPS handwritten digit dataset. We conclude in Section 6.

2. Setting Up the Framework: Single Component in $\mathbb{R}$

We start the discussion by demonstrating the basis of our approach on the problem of estimating parameters of a single univariate Gaussian distribution $p(x) = N(\mu, \sigma^2)$. We first establish a connection between eigenfunctions and eigenvalues of the convolution operator $\mathcal{G}_p^\omega f(y) = \int_{\mathbb{R}} e^{-\frac{(x-y)^2}{2\omega^2}} f(x)\, p(x) dx$ and the parameters μ and σ^2. We show these parameters can be estimated from sampled data. We will need the following

Proposition 1 *(Refinement of a result in [13]) Let $\beta = 2\sigma^2/\omega^2$ and let $H_i(x)$ be the i-th order Hermite polynomial. Then eigenvalues and eigenfunctions of $\mathcal{G}_p^\omega$ for $i = 0, 1, \cdots$ are given by*

$$\lambda_i = \frac{\sqrt{2}}{(1+\beta+\sqrt{1+2\beta})^{1/2}} \left(\frac{\beta}{1+\beta+\sqrt{1+2\beta}}\right)^i \qquad (1)$$

$$\phi_i(x) = \frac{(1+2\beta)^{1/8}}{\sqrt{2^i i!}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\frac{\sqrt{1+2\beta}-1}{2}\right) \times H_i\left(\left(\frac{1+2\beta}{4}\right)^{\frac{1}{4}} \frac{x-\mu}{\sigma}\right) \qquad (2)$$

Since $H_0(x) = 1$, and putting $C = (1+2\beta)^{1/8}$

$$\phi_0(x) = C \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\frac{\sqrt{1+4\sigma^2/\omega^2}-1}{2}\right) \qquad (3)$$

We observe that that the maximum value of $|\phi_0(x)|$ is taken at the mean of the distribution μ, hence $\mu = \text{argmax}_x |\phi_0(x)|$. We also observe that $\frac{\lambda_1}{\lambda_0} = \frac{2\sigma^2}{\omega^2}\left(1 + \frac{2\sigma^2}{\omega^2} + \sqrt{1+\frac{4\sigma^2}{\omega^2}}\right)^{-1}$. Taking $r = \lambda_1/\lambda_0$, we derive

$$\sigma^2 = \frac{r\omega^2}{(1-r)^2}. \qquad (4)$$

Thus we have established an explicit connection between spectral properties of $\mathcal{G}_p^\omega$ and parameters of $p(x)$. We now present **Algorithm 1** for estimating μ and σ^2 from a sample $x_1, \ldots, x_n$ from $p(x)$.

- **Step 1.** Construct kernel matrix K_n, $(K_n)_{ij} = \frac{1}{n} e^{-\frac{(x_i-x_j)^2}{2\omega^2}}$. K_n serves as the empirical version of the operator $\mathcal{G}_p^\omega$. Compute the top eigenvector v_0 of K_n and the top two eigenvalues $\lambda_0(K_n), \lambda_1(K_n)$.

Actual value	$\mu = 0$	$\sigma = 1$
SE ($\hat{\mu}$, $\hat{\sigma}$)	0.000 (0.014)	1.005 (0.012)
Std Est ($\bar{x}$, s)	0.002 (0.011)	1.001 (0.007)

Table 2. Average(standard deviation) of spectroscopic estimator SE($\hat{\mu}$, $\hat{\sigma}$) and the standard estimator Std Est($\bar{x}, s$) of 100 simulation run. In each run, estimators are calculated from 1000 *i.i.d* samples of $N(0, 1)$.

- **Step 2.** Construct estimators $\hat{\mu}$ and $\hat{\sigma}^2$ for mean and variance as follows:

$$\hat{\mu} = x_k, \qquad k = \underset{i}{\operatorname{argmax}} |(v_0)_i|$$

$$\hat{\sigma}^2 = \frac{\omega^2 \hat{r}}{(1 - \hat{r})^2},$$

where $\quad \hat{r} = \frac{\lambda_0(K_n)}{\lambda_1(K_n)}$.

These estimators are constructed by substituting top eigenvector of K_n for the top eigenfunction of $\mathcal{G}_p^\omega$ and eigenvalues of K_n for the corresponding eigenvalues of $\mathcal{G}_p^\omega$.

It is well-known (e.g., [6]) that eigenvectors and eigenvalues of K_n approximate and converge to eigenfunctions and eigenvalues of $\mathcal{G}_p^\omega$ at the rate $\frac{1}{\sqrt{n}}$ as $n \to \infty$, which implies consistency of the estimators. The accuracy of $\hat{\mu}$ and $\hat{\sigma}^2$ depends on how well the empirical operator K_n approximates the underlying operator $\mathcal{G}_p$.

The Table 2 reports the average and the standard deviation of our spectroscopic estimators $(\hat{\mu}, \hat{\sigma}^2)$ compared the standard estimators $(\bar{x}, s^2)$ for one hundred repetitions of the simulation. We see that our spectroscopic estimators are comparable to the standard estimators for mean and variance of a single Gaussian.

3. Setting Up the Framework: Single Component in $\mathbb{R}^d$

In this section we extend our framework to estimating a single multivariate Gaussian $p = N(\mu, \Sigma)$ in $\mathbb{R}^d$. Let $\Sigma = \sum_{i=1}^d \sigma_i^2 u_i u_i^t$ be the spectral decomposition of the covariance matrix Σ. As before we put $\mathcal{G}_p^\omega f(x) = \int_{\mathbb{R}^d} e^{-\frac{\|x-y\|^2}{2\omega^2}} f(y)\, p(y) dy$. Since the kernel $e^{-\frac{\|x-y\|^2}{2\omega^2}}$ is invariant under rotations, it follows that the operator $\mathcal{G}_p^\omega$ can be decomposed as: $\mathcal{G}_p^\omega = \oplus_{i=1}^d \mathcal{G}_{p_i}^\omega$, where p_i is an 1-dimensional Gaussian with variance σ_i^2 and mean $\langle \mu, u_i \rangle$ along the direction of u_i.

It is easy to see that given two operators $\mathcal{F}$, $\mathcal{H}$, the spectrum of their direct sum $\mathcal{F} \oplus \mathcal{H}$ consists of pairwise products $\lambda\mu$, where λ and μ are their respective eigenvalues. The corresponding eigenfunction of the product is $e_{[\lambda,\mu]}(x, y) = e_\lambda(x)\, e_\mu(y)$.

Applying this result, we see that eigenvalues and eigenfunctions of of $\mathcal{G}_p^\omega$ can be written as products

$$\lambda_{[i_1,\ldots,i_d]}(\mathcal{G}_p^\omega) = \prod_{j=1}^d \lambda_{\,i_j}(\mathcal{G}_{p_j}^\omega)$$

$$\phi_{\,[i_1,\ldots,i_d]}(\mathcal{G}_p^\omega)(x) = \prod_{j=1}^d \phi_{\,i_j}(\mathcal{G}_{p_j}^\omega)(\langle x, u_j \rangle)$$

Where $[i_1, \ldots, i_d]$ is a multindex over all components. It can be seen that $\phi_{\,[0,\ldots,0]}$ is (up to a scaling factor) a Gaussian with the same mean μ as the original distribution $p(x)$. Thus μ can be estimated as the point with maximum value $\phi_{\,[0,\ldots,0]}$ in the same way as for 1-dimensional distributions.

Consider now ϕ_I, where $I = [\underbrace{0, \ldots, 0}_{i-1}, 1, 0, \ldots, 0]$. Since $H_2(x) = 2x$, Eq. 1 implies that $\frac{\phi_I(x)}{\phi_{\,[0,\ldots,0]}(x)}$ is a linear function in x with the gradient pointing in the direction of u_i. That allows us to estimate the principal directions. The resulting **Algorithm 2** for estimating μ and Σ is presented below:

Step 1. Construct kernel matrix K_n, $(K_n)_{st} = \frac{1}{n} e^{-\frac{\|x_s - x_t\|^2}{2\omega^2}}$. K_n serves as the empirical version of the operator $\mathcal{G}_p^\omega$. Compute eigenvalues $\lambda(K_n)$ and eigenvectors $v(K_n)$ of K_n. Denote the top eigenvector by v_0 and the corresponding eigenvalue by λ_0.

Step 2. Identify each eigenvector v_i, $v_i \neq v_0$, $i = 1, \ldots, d$ such that the values of $\frac{v_i}{v_0}$ are approximately linear in x, that is

$$\frac{v_i(x_s)}{v_0(x_s)} \approx a^T x_s + b, \qquad a, b \in \mathbb{R}^d$$

The corresponding principal direction u_i is estimated by $\hat{u}_i = \frac{a}{\|a\|}$. Let the corresponding eigenvalue be λ_i.

Step 3. Construct estimators $\hat{\mu}$ and $\hat{\Sigma}$ for mean and variance as follows:

$$\hat{\mu} = x_k, \qquad k = \underset{i}{\operatorname{argmax}} |(v_0)_i|$$

$$\hat{\Sigma} = \sum_{i=1}^d \hat{\sigma}_i^2 \hat{u}_i \hat{u}_i^t,$$

where $\hat{\sigma}_i^2 = \frac{\omega^2 \hat{r}_i}{(1-\hat{r}_i)^2} \quad$ and $\quad \hat{r}_i = \frac{\lambda_0}{\lambda_i}$.

4. Spectroscopic Estimation for Mixtures of Gaussians

We now extend our framework to the case of a mixture of several multivariate Gaussian distributions with potentially different covariance matrices and mixture coefficients. To illustrate our approach we sample 1000 points from two different Gaussian distributions $N(2, 1^2)$ and $N(-2, 1^2)$ and from their mixture $0.5N(2, 1^2) + 0.5N(-2, 1^2)$. The histogram of the mixture density is shown in the top left panel of Fig 2, and histograms of each mixture component are shown in the right top panels. Taking the bandwidth $\omega = 0.3$, we construct three kernel matrices K^1, K^2 and K for a sample from each of the components and the mixture distribution respectively. The middle and lower left panels show the top two eigenvectors of K, while the middle and lower right panels show the top eigenvector of K^1 and K^2 respectively.

The key observation is to notice the similarity between the left and right panels. That is, the top eigenvectors of the mixture are nearly identical to the top eigenvectors of each of the components. Thus knowing eigenvectors of the mixture allows us to approximate top eigenvectors (and the corresponding eigenvalues) for each of the components. Having access to these eigenvectors and using our Algorithms 1,2, allows us to estimate parameters of each of the mixture components.

This phenomenon is easily understood from the point of view of operator theory. The leading eigenfunctions of operators defined by each mixture component are approximately the eigenfunctions of the operators defined on the mixture distribution. To be explicit, let us consider the Gaussian convolution operator $\mathcal{G}_p^\omega$ defined by the mixture distribution $p(x) = \pi^1 p^1 + \pi^2 p^2$, with Gaussian components $p^1 = N(\mu^1, \Sigma^2)$ and $p^2 = N(\mu^2, \Sigma^2)$ and the Gaussian kernel $K(x, y)$ with bandwidth ω. The corresponding operators are $\mathcal{G}_{p^1}^\omega$ and $\mathcal{G}_{p^2}^\omega$ and $\mathcal{G}_p^\omega = \pi^1 \mathcal{G}_{p^1}^\omega + \pi^2 \mathcal{G}_{p^2}^\omega$ respectively. Consider an eigenfunction $\phi^1(x)$ of $\mathcal{G}_{p^1}^\omega$ with eigenvalue λ^1, $\mathcal{G}_{p^1}^\omega \phi^1 = \lambda^1 \phi^1$. We have

$$\mathcal{G}_p^\omega \phi^1(y) = \pi^1 \lambda^1 \phi^1(y) + \pi^2 \int K(x, y) \phi^1(x) p^2(x) dx.$$

It can be shown that eigenfunction $\phi^1(x)$ of $\mathcal{G}_{p^1}^\omega$ is centered at μ^1 and decays exponentially away from μ^1. Therefore, assuming the separation $\|\mu^1 - \mu^2\|$ is large enough, the second summand $\pi^2 \int K(x, y)\phi^1(x) p^2(x) dx \approx 0$ for all y uniformly, and hence $\mathcal{G}_p^\omega \phi^1 \approx \pi^1 \lambda^1 \phi^1$. When the approximation holds the top eigenfunctions of $\mathcal{G}_p^\omega$ are approximated by top eigenfunctions of either $\mathcal{G}_{p^1}^\omega$ or $\mathcal{G}_{p^2}^\omega$.

Theorem 1 *Given a d-dimensional mixture of two Gaussians $p(\mathbf{x}) = \sum_{i=1}^2 \pi^i p^i(x)$ where π_i is mixing weight and p_i is the density corresponding to $N(\mu_i, \sigma^2 I)$. Define $\beta = 2\sigma^2/w^2$ and $\xi = \sqrt{2}\sigma/\sqrt{\sqrt{1+2\beta}-1}$, then the first eigenfunction (ϕ_0^1 with an eigenvalue λ_0^1) of $\mathcal{G}_{p_1}^w$ is approximately an eigenfunction of $\mathcal{G}_p^w$ in the following sense: For any $\epsilon > 0$ we have that for all y*

$$\mathcal{G}_p^w \phi_0^1(y) = \pi_1 \lambda_0^1 (\phi_0^1(y) + T(y)) \quad \textit{and} \quad |T(y)| \le \epsilon$$

assuming that the separation satisfies

$$\frac{\|\mu_1 - \mu_2\|^2}{\xi^2 + \sigma^2} \ge 2 \log\left(\frac{\pi_2}{\pi_1}\right) + 2 \log\left(\frac{1}{\epsilon}\right) + \frac{d}{4} \log(1 + 2\beta)$$

We do not provide a proof of Theorem 1 for lack of space. A more general version of the theorem for several Gaussians with different covariance matrices can also be given along the same lines. Together with some perturbation analysis ([5]) it is possible to provide bounds on the resulting eigenvalues and eigenfunctions of the operator.

We now observe that for the operator $\mathcal{G}_{p^g}^\omega$, the top eigenfunction is the only eigenfunction with no sign change. Therefore, such eigenfunction of $\mathcal{G}_p^\omega$ corresponds to exactly one component of the mixture distribution. This immediately suggest a strategy for identifying components of the mixture: we look for eigenfunctions of $\mathcal{G}_p^\omega$ that have *no sign change*. Once these eigenfunctions of $\mathcal{G}_p^\omega$ are identified, each eigenfunction of $\mathcal{G}_p^\omega$ can be assigned to a group determined an eigenfunction with no sign change. As a result, the eigenvalues and eigenfunctions in each group only depend on one of the component p^g and mixing weight π^g. By reversing the relationship between parameters and eigenvalues/eigenfunctions, parameter estimations for each mixing component can be constructed based only on the eigenvalues/eigenvectors in the corresponding group.

4.1. Algorithm for Estimation of a Mixture of Gaussians

Following the discussion above, we now describe the resulting algorithm for estimating a multidimensional mixture of Gaussians $p(x) = \sum_{g=1}^G \pi^g N(\mu^g, \Sigma^g)$, from a sample $x_1, \ldots, x_n \in \mathbb{R}^d$, first giving the following

Definition 1 *For vectors $d, e \in \mathbb{R}^n$), we define*
1. **ϵ-support** *of d is the set of indices $\{i$: $|d_i| \ge \epsilon$, $i = 1, \cdots, n\}$.*
2. *d has* **no sign changes up to precision** *ϵ, if d is*

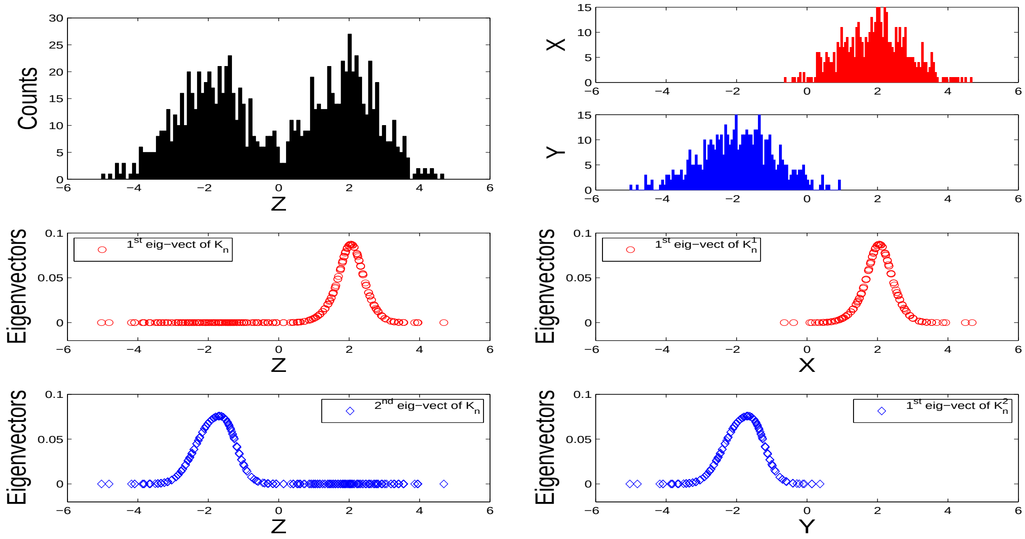

Figure 2. Eigenvectors of a Gaussian kernel matrix ($\omega = 0.3$) of 1000 data sampled from a Mixture Gaussian distribution $P = 0.5N(2, 12) + 0.5N(-2, 12)$. Top left panel: Histogram of the data. Middle left panel: First eigenvector of K_n. Bottom left panel: Second eigenvector of K_n. Top right panel: Histograms of data from each component. Middle right panel: First eigenvector of K_n^1. Bottom right panel: First eigenvector of K_n^2.

either positive or negative on the ϵ-support of e. $\{i : |e_i| \geq \epsilon\} \subset \{i : |d_i| \geq \epsilon\}$.

Algorithm 3. Spectroscopic estimation of a Gaussian mixture distribution.

Input: Data $x_1, \ldots, x_n \in \mathbb{R}^d$. **Parameters:** Kernel bandwidth $\omega > 0$, threshold $\epsilon > 0$.[1]
Output: Number of components $\hat{G}$. Estimated mean $\hat{\mu}^g \in \mathbb{R}^d$, mixing weight $\hat{\pi}^g$, $g = 1, \ldots, \hat{G}$ and covariance matrix Σ^g for each component.

- **Step 1.** Constructing K_n, the empirical approximation to $\mathcal{G}_p^\omega$:

  Put $(K_n)_{ij} = \frac{1}{n} \exp\left(-\frac{\|x_i - x_j\|^2}{2\omega^2}\right)$, $i, j = (1, \ldots, n)$. Compute the (leading) eigenvalues $\lambda_1, \lambda_2, \ldots$ and eigenvectors $v_1, v_2, \ldots$ of K_n.

- **Step 2.** Estimating the number of components G:

  Identify all eigenvectors of K_n, which have no sign changes up to precision ϵ. Estimate G by the number ($\hat{G}$) of such eigenvectors and denote those eigenvectors and the corresponding eigenvalues by $v_0^1, v_0^2, \ldots, v_0^{\hat{G}}$ and $\lambda_0^1, \lambda_0^2, \ldots, \lambda_0^{\hat{G}}$ respectively.

- **Step 3.** Estimating the mean μ^g and the mixing weight π^g of each component:

  For the g'th component, $g = 1, \ldots, \hat{G}$, estimate the mean and the mixing weight as follows:

$$\hat{\mu}^g = x_k, \quad \text{where} \quad k = \operatorname*{argmax}_i |(v_0^g)_i|$$

$$\hat{\pi}^g = \frac{n^g}{\sum_{h=1}^{\hat{G}} n^h},$$

  where $\quad n^h = \text{cardinality of } \epsilon\text{-support of } v_0^h.$

To estimate the covariance matrix Σ^g of each component p^g: we first all eigenvectors such that $\frac{v(x_s)}{v_0^g(x_s)}$ is approximately a linear function of x_s on the ϵ-support of v_0^g. Then we can apply the estimation methods described in **Algorithm 2**, **Step 3** on the ϵ-support of v_0^g.

5. Simulations and Experiments

Simulation: multivariate Gaussian mixture distribution.

A simulation on five dimensional data is carried out to test the proposed algorithm. The first two variables X_1 and X_2 are a mixture three Gaussian components $p(X) = \sum_{g=1}^3 \pi^g N(\mu^g, \Sigma^g)$ with mixing weights and group means shown in Table 3 and covariance matrices:

$$\Sigma^1 = \begin{pmatrix} 0.5 & -0.25 \\ -0.25 & 0.5 \end{pmatrix}, \quad \Sigma^2 = \begin{pmatrix} 0.5 & 0.25 \\ 0.25 & 0.5 \end{pmatrix},$$

[1] In our implementation of the algorithm we choose $\epsilon = \max_j |(v_i)_j|/n$ for each eigenvector v_i. In the description of the algorithm we will use the same ϵ for simplicity.

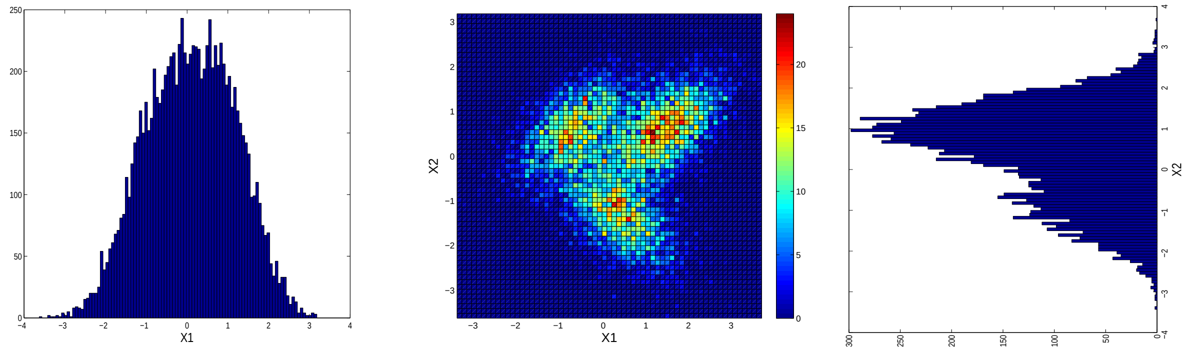

Figure 3. Left: Histogram of the first coordinate X_1; Middle: Two dimensional histogram of the first two coordinates (X_1, X_2). Right: Histogram of X_2.

$$\Sigma^3 = \begin{pmatrix} 0.5 & -0.25 \\ -0.25 & 0.5 \end{pmatrix}$$

The remaining three variables are Gaussian noise $N(0, 0.1I)$. In each simulation run, 3000 data points are sampled. The histogram of X_1, two-dimensional histogram of X_1 and X_2, and histogram of X_2 for one simulation run are shown in Figure 3. We see that it is impossible to identify the number of components by investigating the one-dimensional histograms. The Algorithm 3 with $\omega = 0.1$ was used to estimate the number of components G, mixing weights π^g. The simulation is run 50 times and the algorithm accurately estimated the number of groups in 46 of the 50 runs. Two times the number of groups was estimated as 2 and two times as 4. The average and standard deviation of the estimators of mixing weights and means for the 46 runs are reported in Table 3. We see that the estimates for mixing weights are close to the true values and the estimated group means are close to the estimates from labeled data. Covariance estimates, which we do not show due to space limitations, also show reasonable accuracy.

USPS ZIP code data.
To apply our method to some real-world data we choose a subset of the USPS handwritten digit dataset, consisting of 16x16 grayscale images. In this experiment, 658 “3”s, 652 “4”s, and 556 “5”s in the training data are pooled together as our sample (size 1866). The Spectroscopic estimation algorithm using a Gaussian kernel with bandwidth 2 is applied to the sample . Here we do not use the algorithm to estimate mean and variance of each component, since we do not expect the distribution of the 256 dimensional data to like a Gaussian distribution. Instead, we investigate the eigenvectors with no sign change over $\{x : |v(x)| > \epsilon\}$. We expect (1) the data corresponding to large absolute values of each of such eigenvectors present one mode (cluster) and (2) those data points are in the same digit group.

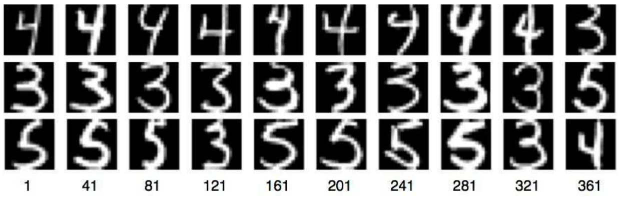

Figure 4. Images ordered by the three eigenvectors v_1, v_{16} and v_{49} identified by Algorithm 3. The images are the digits corresponding to the 1^{st}, 41^{st}, 81^{st}, $\cdots$, 361^{st} largest entries of $|v_1|$ (first row), $|v_{16}|$ (second row) and $|v_{49}|$ (third row).

	“3” (T)	“4” (T)	“5” (T)
“3” (P)	625	0	45
“4” (P)	17	640	32
“5” (P)	16	12	479

Table 4. Confusion matrix of clustering results for USPS handwritten digits. Each cell shows the number of data points belonging both in the **T**rue group (e.g. “3”) and the **P**redicted group (e.g. “3”)

In the output of our algorithm, three eigenvectors v_1, v_{16} and v_{49} of K_n satisfy the condition of no sign change over $\{x : |v(x)| > \epsilon\}$ with $\epsilon = \max(v)/n$. We first rank the data by an decreasing order of $|v|$ and show the 1^{st}, 41^{st}, 81^{st}, $\cdots$, 361^{st} digits in Figure 4. All digits with larger value of $|v_1|$ belong to the group of “4”s, and other digits (“3” and “5”) correspond to smaller values of $|v_1|$. Similarly, larger values of $|v_{16}|$ are in the group of “3”s and $|v_{49}|$ for “5”s.

By assigning digits to their component defined by one of the eigenvectors (v_1, v_{16}, v_{49}) we obtain the clustering results shown in the confusion Table 4. We see that

Parameter value	$\pi^1 = 0.4$	$\pi^2 = 0.3$	$\pi^3 = 0.3$
Spectroscopy (STD)	0.40 (0.03)	0.30 (0.03)	0.30 (0.03)

Parameter value	$\mu_1^1 = 1$	$\mu_2^1 = 1$	$\mu_1^2 = 0$	$\mu_2^2 = -1$	$\mu_1^3 = -1$	$\mu_2^3 = 1$
Spectroscopy (STD)	1.00 (0.12)	1.00 (0.19)	0.01 (0.20)	-0.94 (0.21)	-0.96 (0.22)	0.99 (0.22)
$\bar{x}(STD)$ of each group	1.00 (0.02)	1.00 (0.022)	-0.00 (0.03)	-1.00 (0.02)	-1.00 (0.02)	0.99 (0.03)

Table 3. Estimation of mixing weight and mean of each component

the overall accuracy of clustering is 93.46%. This clustering method can be thought of as an extension of the framework provided in this paper. While this method is closely related to spectral clustering, the procedures for choosing eigenvectors are different.

6. Conclusion

In this paper we have presented *Data Spectroscopy*, a new framework for inferring parameters of certain families of probability distributions from data. In particular we have analyzed the case of a mixture of Gaussian distributions and shown how to detect and estimate its components under the assumption of reasonable component separation. The framework is based on the spectral properties of data-dependent convolution operators and extends intuitions from spectral clustering and Kernel PCA. We have developed algorithms and have shown promising experimental results on simulated and real-world datasets.

We think that our approach provides new connections between spectral methods and inference of distributions from data, which may lead to development of algorithms for using labeled and unlabeled data in problems of machine learning.

7. Acknowledgments

The authors thank Yoonkyung Lee for helpful discussions and suggestions. Mikhail Belkin was partially supported by NSF Early Career Award 0643916. Bin Yu was partially supported by NSF grant DMS-0605165, ARO grant W911NF-05-1-0104, NSFC grant 60628102, a grant from MSRA, and a Guggenheim Fellowship in 2006.

References

[1] M. Belkin, P. Niyogi, *Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering*, NIPS 2001.

[2] S. Dasgupta, *Learning mixtures of Gaussians*, FOCS 1999.

[3] A. Dempster, N. Laird, D. Rubin, *Maximum-likelihood from incomplete data via the em algorithm*, Journal of Royal Statistics Society, Ser. B, 39 (1997), pp. 1–38.

[4] R.Kannan, H.Salmasian, S.Vempala, *The spectral method for general mixture models*, COLT 2005.

[5] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, 1966.

[6] V. Koltchinskii and E. Giné, *Random matrix approximation of spectra of integral operators*, Bernoulli, 6 (2000), pp. 113 – 167.

[7] S. MacEachern and P. Muller, *Estimating mixture of Dirichlet process models*, Journal of Computational and Graphical Statistics, 7 (1998), pp. 223–238.

[8] A. Ng, M. Jordan, and Y. Weiss, *On spectral clustering: Analysis and an algorithm*, NIPS 2001.

[9] K. Pearson, *Contributions to the mathematical theory of evolution*, Phil. Trans. Royal Soc., 185A, (1894), pp. 71–110.

[10] R. Redner and H. Walker, *Mixture densities, maximum likelihood and the em algorithm*, SIAM Review, 26 (1984), pp. 195–239.

[11] B. Schölkopf, A. J. Smola, and K.-R. Müller, *Kernel principal component analysis*, Advances in kernel methods: support vector learning, (1999), pp. 327–352.

[12] C.Williams, M.Seeger, *The effect of the input density distribution on kernel-based classifiers*, ICML2000.

[13] H. Zhu, C. Williams, R. Rohwer, and M. Morcinie, *Gaussian regression and optimal finite dimensional linear models*, in Neural networks and machine learning, C. Bishop, ed., 1998.